Jinkwon Kim

School of Computing, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
coco@kaist.ac.kr +82-10-2754-9153 https://sites.google.com/view/jinkwonkim

RESEARCH INTERESTING

My primary interest lies in cross-layer optimizations for efficient compression-based systems. Due to the exponential growth of data utilized and generated by key workloads (e.g., scientific computing, machine learning, and graph analytics), compression-based systems have become indispensable in various hardware components. I have optimized several different compression-based systems through cross-layer optimizations (e.g., CPU ISA [*TC'20*], main memory [*HPCA'22*], DNN [*DATE'22, TC (under review)*], SSD [*TC'22*], and sparse tensor accelerator [*MICRO'23*]). Currently, I am researching to further enhance the hardware-based pseudo-tiling proposed in *MICRO'23*.

EDUCATION

KAIST, Daejeon, South Korea — Mar 2017 – Feb 2024 *(expected)*

- Integrated M.S./Ph.D Program, Computer Science
- Advisor: Soontae Kim
- GPA: 3.98 / 4.3

Hanyang University, Seoul, South Korea — Mar 2012 – Aug 2016

- Bachelor, Double Majors in Industrial Engineering / Computer Science and Engineering
- GPA: 4.11 / 4.5

PUBLICATIONS

[1] **Jinkwon Kim**, Myeongjae Jang, Haejin Nam, and Soontae Kim, "HARP: Hardware-Based Pseudo-Tiling for Sparse Matrix Multiplication Accelerator", accepted in *IEEE/ACM International Symposium on Microarchitecture (**MICRO**)*, 2023.

[2] Myeongjae Jang, **Jinkwon Kim**, Haejin Nam, and Soontae Kim, "Zero and Narrow-width Value-aware Compression for Quantized Convolutional Neural Networks", accepted in *IEEE Transactions on Computers (**TC**)*.

[3] Mincheol Kang, Wonyoung Lee, **Jinkwon Kim**, and Soontae Kim, "PR-SSD: Maximizing Partial Read Potential by Exploiting Compression and Channel-Level Parallelism", *IEEE Transactions on Computers (**TC**)*, Vol.72, No.3, pp.772-785, May 2022.

[4] Myeongjae Jang, **Jinkwon Kim**, Jesung Kim, and Soontae Kim, "ENCORE Compression: Exploiting Narrow-width Values for Quantized Deep Neural Networks", *Design, Automation, and Test in Europe (**DATE**)*, Antwerp, Belgium, Mar 2022.

[5] **Jinkwon Kim**, Mincheol Kang, Jeongkyu Hong, and Soontae Kim, "Exploiting Inter-block Entropy to Enhance the Compressibility of Blocks with Diverse Data", *IEEE International Symposium on High-Performance Computer Architecture (**HPCA**)*, Seoul, South Korea, Apr 2022.

[6] **Jinkwon Kim**, Seokin Hong, Jeongkyu Hong, and Soontae Kim, "CID: Co-Architecting Instruction Cache and Decompression System for Embedded Systems", *IEEE Transactions on Computers (**TC**)*, Vol.70, No.7, pp.1132-1145, Jul 2021.

RESEARCH EXPERIENCE

Hardware-based Pseudo-Tiling for Sparse Matrix Multiplication Accelerator

- Redefine the boundary between hardware and software for tiling in sparse matrix multiplication.
- Identify the limitations of the software-based tiling: manual execution, generation of several compression formats for each tile, and ineffectual accesses.
- Introduce a hardware-based pseudo-tiling, which performs the tiling process in hardware instead of software to overcome the aforementioned limitations of the software-based tiling.
- The hardware-based pseudo-tiling allows logical tiling of the original compressed matrix without generating a compression format for each tile and skips ineffectual accesses for input matrices.
- Accepted in MICRO 2023.

Maximizing Partial Read Potential by Exploiting Compression and Channel-Level Parallelism

- Propose a new compression algorithm for applying partial read operations in SSD and a split module that can use partial read operation for uncompressed requests via channel-level parallelism in SSD.
- Published in TC 2022.

Exploiting Narrow-Width Values to Reduce Data Traffic in Quantized Deep Neural Networks

- Propose a new compression algorithm based on the narrow-width value property in modern quantized DNN to reduce data traffic.

- Published in DATE 2022 and accepted in TC 2023

Exploiting Inter-Block Entropy to Improve the Compressibility of Blocks with Diverse Data

- Leverage data patterns in software to overcome the limitations of previous intra- and inter-block compression techniques.
- Discover the natural low-entropy among blocks and propose three optimization techniques to generate artificial low-entropy among blocks. Based on these two low-entropy types, we propose an entropy-based inter-block pattern compression technique.
- Propose hardware-based and profiling-based pattern selection methods for efficient pattern management.
- Propose a hybrid approach that leverages both intra- and inter-block compression techniques.
- Published in HPCA 2022.

Co-Design Compression-Support Architecture and Code Compression for Low-Power and Low-Area

- Leverage software-layer characteristics to optimize the code compression technique and the hardware components.
- Discover that certain bits within the 32-bit instruction encoding in RISC ISAs have high entropy due to several characteristics of high-level languages, such as reusability and the calling convention.
- Based on these observations, we re-organize the hardware components of the code compression-support architecture and design the instruction cache architecture to efficiently support the proposed code compression technique.
- Published in TC 2020.

AWARDS & HONORS

▪ National Scholarship, KAIST	2017 – present
▪ Summa Cum Laude, Hanyang University	2016

SKILLS

- **Programmings:** C/C++, Python, Verilog, Chisel
- **Architecture Simulators and Tools:** Gem5, ZSim, SST, Pin, Synopsys Design Compiler, Ramulator, DRAMSim2, DRAMPower, McPAT, CACTI
- **System Software:** Linux, Warewulf HPC Cluster, QEMU

TEACHING EXPERIENCE

▪ Teaching Assistant for Digital System and Lab, KAIST	Spring 2020
▪ Teaching Assistant for Computer Architecture, KAIST	Spring 2019
▪ Teaching Assistant for Computer Organization, KAIST	Fall 2018
▪ Teaching Assistant for Computer Organization, KAIST	Spring 2018
▪ Teaching Assistant for Computer Organization, KAIST	Fall 2018